Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability

Mireia Yurrita
m.yurritasemperena@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Tim Draws
t.a.draws@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Agathe Balayn
a.m.a.balayn@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Dave Murray-Rust
D.S.Murray-Rust@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Nava Tintarev
n.tintarev@maastrichtuniversity.nl
Maastricht University
Maastricht, The Netherlands

Alessandro Bozzon
A.Bozzon@tudelft.nl
Delft University of Technology
Delft, The Netherlands

ABSTRACT

Recent research claims that information cues and system attributes of algorithmic decision-making processes affect decision subjects' fairness perceptions. However, little is still known about how these factors interact. This paper presents a user study ($N = 267$) investigating the individual and combined effects of explanations, human oversight, and contestability on informational and procedural fairness perceptions for high- and low-stakes decisions in a loan approval scenario. We find that explanations and contestability contribute to informational and procedural fairness perceptions, respectively, but we find no evidence for an effect of human oversight. Our results further show that both informational and procedural fairness perceptions contribute positively to overall fairness perceptions but we do not find an interaction effect between them. A qualitative analysis exposes tensions between information overload and understanding, human involvement and timely decision-making, and accounting for personal circumstances while maintaining procedural consistency. Our results have important design implications for algorithmic decision-making processes that meet decision subjects' standards of justice.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Collaborative and social computing**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

explanations, human oversight, contestability, fairness perceptions, algorithmic decision-making

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581161

ACM Reference Format:
Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3544548.3581161

1 INTRODUCTION

Motivated by concerns about bias and discrimination in algorithmic decision-making [73], recent work has developed fairness-aware algorithmic systems [6, 33, 108] that ensure outcome distribution equity [32, 42]. However, even when a decision-making process is fair by some objective standard, decision subjects might not *perceive* it as fair [59] if aspects such as the inscrutability and unaccountability often surrounding algorithmic systems [17] go against their standards of justice [58, 72, 96].[1] Perceptions of unfairness could, in turn, jeopardize end users' trust in normatively fair algorithmic decision-making processes and, therefore, be an obstacle for their broader acceptance [31, 58, 72, 96, 103]. That is why a growing body of human-computer interaction (HCI) literature now focuses on determining which factors – e.g., information cues [63] such as explanations [17, 30, 71, 83] and system attributes [63] such as human oversight[2] [29, 65, 66, 70, 103] or contestability [68, 93] – effectively contribute to decision subjects' fairness perceptions.

Despite making important contributions, previous HCI research investigating fairness perceptions in algorithmic decision-making has faced two important limitations. First, earlier work has largely studied information cues and system attributes in isolation (e.g., [68, 93]). Such an approach fails to consider the entangled nature of

[1] According to Cropanzano [27], *justice* is a multi-dimensional construct that studies *fairness* perceptions across each of its dimensions. For instance, *procedural justice* refers to a justice dimension that aims to capture fairness perceptions regarding the *process* of a decision (i.e., *procedural fairness perceptions*). Colquitt and Rodell [25] refer to *faceted fairness* as measurements of appropriateness that evoke different justice dimensions.

[2] Throughout this paper, *human oversight* refers to a configuration where human intelligence is applied to identify and correct potential mistakes made by an algorithmic system [5]. We also call this configuration a *hybrid* human-artificial intelligence (AI) decision-making process.

these cues and attributes and does not align with the scenarios contemplated by regulatory efforts such as the European Union's *General Data Protection Regulation* (GDPR) [101]. For example, decision subjects can only meaningfully exercise their *right to contest* an algorithmic decision when they have solid arguments, which require explanations of the decision-making process [79, 101]. Contestation mechanisms and explanations thus co-shape the procedural justice principle of correctability [62] and may, therefore, co-mediate decision subjects' perceptions of procedural fairness [40, 62]. Not considering these entanglements could lead to blind spots regarding how different factors that are theoretically claimed to affect fairness perceptions (e.g., [93]) actually contribute to these perceptions.

Second, prior work has mainly used one-dimensional approaches for measuring fairness perceptions [9, 30, 58, 68, 71, 74, 80, 103, 110]. Although measuring such *overall fairness perceptions* is useful for capturing a global perception of appropriateness [25], prior work on legal and organizational psychology has often advocated for capturing fairness perceptions across up to four different dimensions (i.e., *faceted fairness perceptions*) [21, 24]. These dimensions include perceptions towards the equitable allocation of outcomes (i.e., *distributive fairness perceptions*) [1, 28], the nature of the process that leads to those decisions (i.e., *procedural fairness perceptions*) [62, 64, 89] as well as the information (i.e., *informational fairness perceptions*) [15, 40, 85] and the treatment (i.e., *interpersonal fairness perceptions*) [15] received by decision subjects. Capturing how dimension-specific fairness perceptions manifest may help identify problematic aspects of algorithmic configurations. Additionally, learning how these dimension-specific fairness perceptions combine could then inform the prediction of global perceptions of appropriateness [25]. We argue that prioritizing the measurement of overall fairness might impede the development of a nuanced understanding of how different factors contribute to different facets of users' fairness perceptions [24].

This paper takes a first step towards a nuanced understanding of how different information cues (i.e., explanations) and system attributes (i.e., human oversight and contestability) co-mediate multi-dimensional (i.e., informational and procedural) perceptions of fairness. Given the task-dependent nature of fairness perceptions [9, 17, 58, 70, 86, 96], we account for the stakes of the task as an additional contextual factor. Three research questions guide our work:

- **RQ1:** Do explanations, human oversight, and contestability affect perceived informational and procedural fairness in algorithmic decision-making processes?
- **RQ2:** Do the stakes (high/low) involved in the decision have an effect on perceived informational and procedural fairness?
- **RQ3:** Do users' perceived informational and procedural fairness predict overall perceived fairness?

To address these research questions, we first conducted a preliminary study to surface the interplay between explanations, human oversight and contestability (Section 4.1). We then used these findings to design an online, preregistered[3] user study where participants were shown a fictional loan approval process (Section 4.2). The descriptions shown to participants included information about the decision-making process with or without explanations, with or without human oversight and with or without the right to contest the decision (**RQ1**). Each participant was randomly assigned to a low-stakes[4] (holiday) or to a high-stakes (home) loan approval scenario (**RQ2**). For each scenario, we measured perceptions of informational, procedural and overall fairness (**RQ3**).

Our results show that explanations and contestability affect end users' informational[5] and procedural fairness perceptions, respectively (**RQ1**; see Section 5.2). We do not find evidence that end users' perceptions of informational and procedural fairness are influenced by human oversight (**RQ1**) or the stakes of the task (**RQ2**). Our results further show that perceptions of informational and procedural fairness both relate positively to perceptions of overall fairness, but we do not find an interaction effect between them (**RQ3**). As part of our exploratory analyses, we unpack informational and procedural fairness perceptions into the sub-elements that compose each dimension (Section 5.3). We find that end users may rate perceptions of procedural voice and outcome influence negatively, even when contestability (in the form of appeal processes) is incorporated. We also find that including human oversight may deteriorate perceptions of process consistency and lack of bias. Through a qualitative analysis, we identify three areas of tension: (1) amount of information vs. generating understanding for all, (2) human involvement vs. timely decision-making, and (3) standardized fact-based process vs. accounting for personal circumstances (see Section 5.4). These insights set the grounds for motivating the exploration of transparency beyond outcome explanations, for crafting alternative human-AI configurations, and for designing contestation mechanisms that effectively give voice to decision subjects.

Supplementary materials linked to this paper include task design, preregistration, data, and code for statistical analysis and are openly available at https://osf.io/zrfty/.

2 RELATED WORK

This section describes previous research on how explanations, human oversight, and contestability contribute to fairness perceptions in algorithmic decision-making and discusses the task-dependent nature of this work. We focus on these specific information cues and system attributes as they are directly addressed by Article 22(3) of the GDPR [101]. We then cover research on human decision-making, where fairness perceptions have been captured across multiple dimensions.

2.1 Factors Affecting Perceptions of Fairness in Algorithmic Decision-Making

Explanations. Explanations (i.e., representations of a system's ability to account for their own operation in ways that help users understand how these tasks are being accomplished [17]) are considered key elements for enhancing users' fairness perceptions in algorithmic decision-making processes. Previous work has demonstrated the positive effect of different explanation styles on decision subjects' feelings of justice [17, 30] and their confidence in the fairness of algorithmic systems [72]. Schoeffer et al. [83] found that

[3]The preregistration is openly available at https://osf.io/4uf3m.

[4]Loan approval decisions are generally seen as high-stakes [26] but we still expect differences in users' perceived stakes depending on the loan purpose.

[5]This result replicates and confirms a finding from earlier work [83].

the amount of information in explanations was positively related to *informational* fairness perceptions.

Human Oversight. The term *human oversight* has been used to refer to the configuration where human intelligence is applied to identify potential mistakes in algorithmic decision-making processes [5]. Since algorithmic systems can perform increasingly complex tasks [106], recent research has pointed to opportunities for crafting more reliable and timely decision-making processes with human-artificial intelligence (AI) collaborations [12, 109]. Despite this growing interest, most recent work on fairness perceptions has focused on comparing algorithmic systems with their human counterparts [9, 20, 29, 36, 55, 58, 65, 74] rather than comparing fully automated with hybrid configurations. In one study that did compare algorithmic decision-making to hybrid and human decision-making, Nagtegaal [70] found that hybrid configurations can increase public employees' (subjects of managerial decisions) perceptions of procedural fairness. Wang et al. [103] also evaluated the effect of hybrid decision-making processes on decision subjects' perceptions of fairness but did not find any evidence that hybrid decision-making processes are perceived to be fairer than fully automated ones.

Contestability. Contesting a decision has been defined as the act of opposing an action; either because the action is perceived as mistaken or simply wrong [4, 99]. *Contestability* has, thus, been conceptualized as recourse [48, 91, 98], appeal [99], and as a design principle (i.e., *contestability by design*) [3, 5, 79]. Contestability is said to "surface values" [92] and to be a "form of procedural justice, a way of giving voice to decision subjects, which increases perceptions of fairness" [3]. To the best of our knowledge, however, the effect of contestability in algorithmic decision-making has not yet been widely studied. In one of the few studies that empirically tested the effect of appeals on decision subjects' perceptions of fairness, Vaccaro et al. [93] found that none of their appeal designs improved these perceptions.

Task stakes. Perceptions towards algorithmic decision-making can vary across scenarios [17, 96], based on task characteristics [58], and the stakes of the task (i.e., the impact that a negative outcome would have on the future of an individual [49]) [9, 70, 86]. For instance, Binns et al. [17] found that scenario effects obscure explanation effects under repeated exposure of one explanation style. Lee [58] saw differences in fairness perceptions towards human and algorithmic decision-makers based on task characteristics. Araujo et al. [9] argued that users may perceive algorithmic systems as fairer than human experts only for high-impact decisions in the justice and health domains.

2.2 Capturing Perceptions of Fairness in Decision-Making Processes

Users' perceptions of fairness can be complicated and nuanced [103]. To measure these perceptions in a granular way, disciplines in social sciences such as legal and organizational psychology have empirically validated models that capture perceptions of fairness across different dimensions [25, 27]. These dimensions include perceptions of fairness towards decision outcomes (i.e., *distributive fairness perceptions*) [1, 28], the processes that led to those outcomes (i.e., *procedural fairness perceptions*) [62, 64, 89], the treatment received by decision subjects (i.e., *interpersonal fairness perceptions*) [15], and the information given to decision subjects (i.e., *informational fairness perceptions*) [15, 40, 85]. Each of these dimensions evokes different justice principles and is built upon criteria that have been found to be relevant for that dimension [95]. For instance, procedural fairness perceptions are measured considering perceptions of *procedural voice*, *outcome control*, *consistency of procedures across participants*, *suppression of bias*, *accuracy of factors*, *correctability of outcomes*, and *ethicality of the process* [62, 89].

2.3 Research Gap and Motivation

Although earlier work has shed some light on how to go from a normative to a behavioral understanding of fairness, evidence on how factors that are theoretically related to certain principles of justice co-mediate decision subjects' perceptions of fairness in algorithmic decision-making is still lacking. One reason for this is that the effects of factors believed to enhance perceptions of fairness have been obscured by phenomena such as the *outcome favorability bias* (i.e., divergence in users' perceived fairness based on the favorability of the outcome they receive personally) [74, 103]. For example, although including human oversight has been claimed to bring together the best of the manual and the automatic worlds, there is still little insight into how human oversight contributes to end users' perceptions of fairness. Similarly, although contestability has been claimed to be a key aspect to enhance perceptions of fairness, to the best of our knowledge, there is currently no empirical evidence on whether or how contestability contributes to these perceptions. One could argue that Lyons et al. [67] looked into different modalities of appeal processes and evaluated perceptions of fairness in each case. However, evaluating perceptions of fairness towards different types of appeals is different from evaluating perceptions of fairness towards an algorithmic decision-making process that offers the right to appeal. Another key limitation of previous research is that it did not consider the entangled nature of explanations, human oversight, and contestability. Although decision subjects' right to explanation is not explicitly guaranteed by the GDPR [84], Article 22(3) does explicitly guarantee their right to contest a negative decision [101], for which decision subjects need meaningful (i.e., functional [84]) explanations [79]. The GDPR also states that contestations might vary based on the human intervention in the original decision [101]. Therefore, the way in which a decision can be meaningfully contested depends on the received explanations [79] as well as the interpretation of the implemented safeguards (i.e., right to human intervention, right to express views, and right to contest the decision) [101].

From a methodological perspective, a majority of previous studies has used mono-dimensional (i.e., overall fairness perceptions [25]) approaches for capturing the effects of explanations, human oversight, and contestability on fairness perceptions [9, 30, 58, 68, 71, 74, 80, 103, 110]. This has resulted in a lack of nuance in the understanding of how fairness perceptions are co-mediated by each of these factors. We echo the need to include lessons from the replication crisis within psychology [18] and advocate for a multi-dimensional approach to measuring perceptions of fairness (i.e., faceted fairness perceptions [25]). Although these dimensions were suggested for

human decision-making, we argue that they represent a good starting point toward developing standardized methods for specifically evaluating algorithmic decision-making processes. The benefits of using a more nuanced approach for measuring the effect of explanations on perceptions of fairness have already become evident. Schoeffer et al. [83] found that outcome explanations would increase end users' perceptions of *informational* fairness, but it would make them question structural aspects of the *procedure*, just as it was claimed by Greenberg [40] for human decision-making.

In this paper, we address the above gaps by systematically evaluating algorithmic decision-making processes with varying levels of explanations, human oversight, and contestability, and unpack and disentangle their effects on perceptions of fairness through a multi-dimensional approach. Since the factors (i.e., explanations, human oversight, contestability) that we manipulate in our experimental setting have been related to perceptions of informational and procedural fairness in human decision-making [62, 85], we capture perceptions of fairness across those two dimensions. We also test the predictive validity [24] for these multi-dimensional fairness perceptions on overall fairness perceptions. This enables us to compare the multi-dimensional approach with previously used mono-dimensional approaches.

3 HYPOTHESES

Drawing from literature in legal and organizational psychology for human decision-making [8, 13, 15, 16, 38, 40, 90] and studies on perceptions of fairness in algorithmic systems [9, 41, 58, 72, 80, 83, 93, 96, 103], we formulated eleven hypotheses (Figure 1). Each hypothesis is related to one of the research questions outlined in Section 1 and is followed by a rationale. We preregistered all hypotheses before data collection.

3.1 Hypotheses related to RQ1: Explanations, Human Oversight, and Contestability

- **Hypothesis 1a** (H_{1a}). End users perceive algorithmic decision-making processes as more informationally fair when they are accompanied with explanations.
  Rationale. We extend Schoeffer et al. [83]'s study to evaluate the effect of explanations on informational fairness in both high-stakes and low-stakes decisions. We expect to replicate their findings in our own experimental setting.
- **Hypothesis 1b** (H_{1b}). End users perceive algorithmic decision-making processes as more procedurally fair when these processes are supplemented by human oversight rather than fully automated.
  Rationale. Previous studies have found that users consider human decisions to be fairer than fully automated, algorithmic decisions; especially for practices that are highly complex and are perceived to require human skills [58, 70]. Although recent research has found contradictory evidence on whether users perceive *hybrid* decision-making as fairer than entirely algorithmic decision-making [70, 103], we do expect that human oversight will lead to increased *procedural fairness* perceptions among users in sensitive contexts (e.g., loan approval processes).
- **Hypothesis 1c** (H_{1c}). End users' procedural fairness perceptions differ based on the contestation procedure of an algorithmic decision-making process.
  Rationale. We hypothesize that, as with human decision-making [89], contestation procedures in algorithmic decision-making processes affect perceived procedural fairness.
- **Hypothesis 1d** (H_{1d}). The effect of contestability on end users' procedural fairness perceptions is moderated by the presence of explanations.
  Rationale. Schoeffer et al. [83] found that, although including more information in explanations led to an increased perception of informational fairness, the presence of explanations allowed end users to question the way in which different factors were being used for decision-making. We thus hypothesize that, aside from a general effect of contestability on users' procedural fairness perception (see H_{1c}), the presence of explanations and contestability on the algorithmic decision *interact* in affecting users' perceived procedural fairness.
- **Hypothesis 1e** (H_{1e}). The effect of contestability on end users' procedural fairness perceptions is moderated by the presence of human oversight.
  Rationale. Various studies have demonstrated end users' concern for fully automated, highly complex decision-making processes [58, 70]. That is why we expect that configurations where end users can contest an algorithmic decision lead to varying degrees of procedural fairness perceptions in users depending on whether the original decision was made by a fully automated or hybrid system.

3.2 Hypothesis related to RQ2: Task stakes

- **Hypothesis 2a** (H_{2a}). The effect of explanations on end users' informational fairness perceptions is moderated by the stakes of the task.
  Rationale. Binns et al. [17] found that the nature of the presented scenario moderates the effect of explanation types on fairness perceptions. In line with these findings, we hypothesize that, based on the nature of the task at stake (i.e., involving high or low stakes), end users will be satisfied differently with the amount of information they received.
- **Hypothesis 2b** (H_{2b}). The effect of human oversight on end users' procedural fairness perceptions is moderated by the stakes of the task.
  Rationale. Lee [58] demonstrated that fairness perceptions regarding the decision maker (i.e., a fully automated system or a human) were moderated by task characteristics. Nagtegaal [70] also found that the effect of involving humans on perceptions of procedural justice varied based on the complexity of the task. Despite the context being different (both these studies focused on managerial decisions) and our study considering fully automated vs hybrid decision making, we hypothesize that the stakes of the task (i.e., involving high or low stakes) will similarly moderate the effect of human oversight on procedural fairness perceptions in our study.
- **Hypothesis 2c** (H_{2c}). The effect of contestability on end users' procedural fairness perceptions is moderated by the stakes of the task.

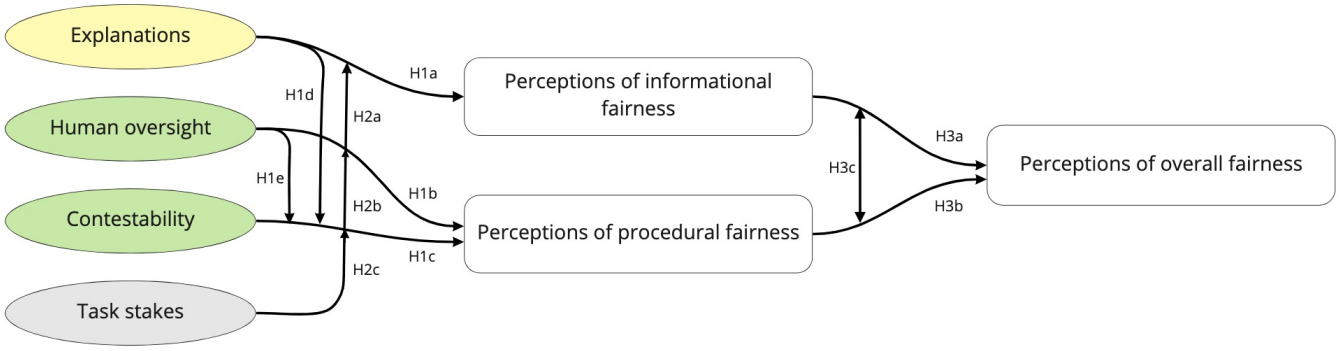

Figure 1: Overview of the hypotheses. Yellow refers to information cues, green to system attributes, and grey to contextual factors.

Rationale. Previous work has suggested that perceptions of fairness regarding the decision-maker generally depend on the nature of the task [58]. We thus hypothesize that the stakes of the task (i.e., involving high or low stakes) also moderate the effect of contestability (e.g., when users are given the right to contest the decision-maker [68]) on users' procedural fairness perceptions.

3.3 Hypothesis related to RQ3: Overall vs. Faceted fairness

- **Hypothesis 3a** (H_{3a}). End users' informational fairness perceptions are positively associated with their overall fairness perceptions.
  Rationale. This hypothesis is in line with findings in human decision-making, where informational fairness was claimed to influence perceptions of overall fairness [24, 39].
- **Hypothesis 3b** (H_{3b}). End users' procedural fairness perceptions are positively associated with their overall fairness perceptions.
  Rationale. Studies dealing with procedural fairness in human decision-making processes [39, 89] demonstrated that participants with a strong influence over the decision-making process were more likely to perceive a negative outcome as fair [47]. We hypothesize that for algorithmic decision-making processes, there will also be a positive relation between perceptions of procedural fairness and overall fairness.
- **Hypothesis 3c** (H_{3c}). End users' perceived informational and procedural fairness interact in predicting overall fairness.
  Rationale. Research in human decision-making has demonstrated that explanations provide the "information needed to evaluate structural aspects of decision-making" [40]. In line with these findings, we hypothesize that perceptions of overall fairness are not just dependent on both informational and procedural fairness, but that these two factors *interact* in predicting overall fairness perceptions.

4 STUDY DESIGN

Because explanations, human oversight, and contestability are entangled by nature [101], we first conducted a preliminary study to craft an experimental setting that would surface the interplay between these factors (Section 4.1). In this exploratory study, we captured preferences towards different explanation styles and investigated what aspects participants would like to contest. We then combined these insights with previous literature to design our main user study in the context of a loan approval process (Section 4.2).

4.1 Preliminary Study

This preliminary study (N = 58) aimed at crafting (1) understandable and (2) actionable[6] explanations that (3) support contestability [101]. We also sought to understand what aspects of the decision-making process participants may contest. Although prior work has already studied the understandability of different types of explanations [17, 30] and identified actionable factors for loan approval processes [83], the interplay between explanations and contestability still represents an underexplored area,[7] hence the need to perform this preliminary study. The design of our preliminary study and the instruments we used to capture participants' preferences can be found in our repository.

4.1.1 Method of the Preliminary Study. As part of our preliminary study, we provided each participant with five types of explanations (randomized) for a fictional home loan denial scenario: (1) *factor importance-based* explanations (i.e., feature importance hierarchy using ">" for expressing "more important than" [83]), (2) *input influence-based* [8] explanations (i.e., list of input variables along with a quantitative measure of the effect and directionality —positive or negative— that each of these variable had on the final decision [17, 30]), (3) *case-based* explanations (i.e., instance from the model's training data that is most similar to the decision being

[6] We define "actionable" factors as the set of variables upon which interventions are possible. We include those variables that may change as a consequence of a change to its causal ancestors (that other authors have named as "mutable but non-actionable" [51])

[7] Although the interplay between explanations and recourse is increasingly being studied (e.g., [50, 87]), for this preliminary study, we do not limit contestability to recourse and inquire whether participants would question other aspects of the decision-making process.

[8] As opposed to some previous work [17, 30], where the quantitative measurement of the input influence was indicated through a varying number of "+" (positive influence) or "-" (negative influence) signs, we expressed this difference in influence through numerical values. We clarified that the number in brackets indicated the magnitude of the positive or negative effect that the variable had on the final decision —negative meaning a contribution towards the rejection decision—.

explained [17, 30]), (4) *counterfactual* explanations (i.e., representation of the alterations that input variables would need for the undesired model output to change [17, 30, 101]), and a combination of (5) *input influence-based and counterfactual* explanations [83]. They were then asked to select the two most understandable and actionable explanations and two explanations thanks to which the decision subject would best know what information to use to contest the decision. We also asked them to choose their overall preferred explanation type. At the end of the study, we included two open-ended questions. The first question aimed to disclose the rationales behind users' preferences for different types of explanations. The second question collected answers on what aspects of the decision-making process participants would be willing to contest. For analyzing the responses to the open-ended questions, we performed a reflexive thematic analysis [19]. Our aim was to use the findings from this preliminary study to inform the design of our main user study (Section 4.2).

4.1.2 Insights from the Preliminary Study. The combination of counterfactuals and input influence-based explanations scored highest for all criteria (see Table 1). To better understand these results, we discuss our findings from the qualitative analysis below. We refer to quotes as Q.*i*, where *i* is the index of a specific quote. Appendix A shows all selected quotes.

Preferences towards different types of explanations. In line with findings from Dodge et al. [30], we found that case-based explanations were considered less fair (Q.1, Q.2). Participants generally preferred explanations that contain more information, which is in line with findings from Schoeffer et al. [83] (Q.3). Moreover, participants generally preferred the combination of input influence-based and counterfactual explanations because these included descriptions of the "how" and a justification of the "why" of decisions, as suggested by Sarra [79]. Input influence-based explanations were regarded as faithful descriptions of how each feature contributes to the algorithm's decision-making process (11/58)[9] (Q.4). Despite using numerical values to indicate different degrees of input influence on the final decision, readability was not flagged as an issue for input influence-based explanations by our participants. Counterfactuals were regarded as concise and explicit when directing the attention to features that were relevant to that particular decision (17/58) (Q.5, Q.6).

What to contest. Participants pointed to two main aspects they would like to contest: first, the basis (i.e., the factors) of the decision and their weights (28/58) (Q.7, Q.8) and second, the usage of an AI (10/58). Algorithmic systems were viewed as lacking subjective judgment capabilities for considering individual circumstances (in line with previous studies [20, 58, 70]) (Q.9). Generalization was also considered to be an inappropriate basis for decision-making (Q.10).

[9]We indicate the prevalence of each statement using proportions (a/b), where *a* indicates the number of participants whose response to the open-ended questions was related to the statement in question, and *b* indicates either the number of participants within a condition that we are specifically referring to or the total number of participants in the study (58 for the preliminary study and 267 for the main study).

4.2 Main User Study

In our main user study, we sought to characterize the main and interaction effects of explanations, human oversight, and contestability on perceptions of informational and procedural fairness. We also explored the influence of contextual factors (i.e., the stakes of the task) in this context and captured the relationship between informational and procedural fairness perceptions and perceptions of overall fairness. We had preregistered our hypotheses, research design, and data analysis plan for the main study before data collection.

4.2.1 Independent Variables. In an effort to minimize the effect of *outcome favorability bias* [103], we followed prior research [9, 83, 86] and showed participants in our user study a fictional loan approval scenario involving the fictional character *Kim* as loan requester. The scenario differed depending on four independent variables. Figure 2 gives an overview of the independent variables and Table 5 in Appendix B shows how each independent variable was displayed in practice.

- *Explanations* (categorical, between-subjects). We assigned each participant to one of two configurations:
  (1) No explanation: participants saw what information the fictional loan requester had been asked to provide but not how this information was used.
  (2) With explanation: participants learned the weight each piece of information had in the final decision (*input influence-based explanation*) and the hypothetical scenarios where the loan requester would have been able to have the loan approved (*counterfactuals*). The factors requested by the bank and the given explanations are inspired by prior work [83] and enhanced based on the insights we got from the preliminary study (Section 4.1). We discarded gender and marital status as decision basis because these factors are explicitly protected by law [14]. Note that the *no explanation* configuration in our study is equivalent to the *disclosure of factors* condition defined by Schoeffer et al. [83], and not to the *baseline without further explanations*. The rationale behind this design choice is twofold: first, we argue that the disclosure of these factors is necessary for participants to be able to judge the fairness of the decision basis. Second, Schoeffer et al. [83] found no difference in informational fairness perceptions between the two aforementioned configurations. These explanations were textual to limit presentation complexity [22, 83, 96].
- *Human oversight* (categorical, between-subjects). We randomly assigned each participant to one of two configurations:
  (1) No human oversight: participants were told that the algorithmic decision-making process was fully automated.
  (2) With human oversight: participants were told that the loan approval process combined the usage of an algorithmic system with human expertise. We designed this condition based on one of the human-in-the-loop configurations discussed by Almada [5]. As opposed to some previous work where a human would supervise each decision made by the algorithmic system [103] — the authors did not find any evidence of this configuration affecting fairness perceptions—, in our study human intervention would serve as a quality control against machine failures [5]. We, therefore, used the confidence of the

	Understandable	Actionable	Supports contestability	Overall
Importance-based explanation	23.64%	17.70%	18.35%	12.08%
Input influence-based explanation	17.27%	20.35%	21.10%	15.52%
Case-based explanation	16.36%	8.85%	14.68%	13.79%
Counterfactual explanation	13.64%	15.93%	16.51%	15.52%
Combination counterfactual & input influence-based	**29.09%**	**37.17%**	**29.36%**	**43.10%**

Table 1: Results from our preliminary exploratory study. We evaluated how (1) understandable and (2) actionable different types of explanations were, and to what extent they (3) supported contestability. Column (4) shows participants' overall preferred option.

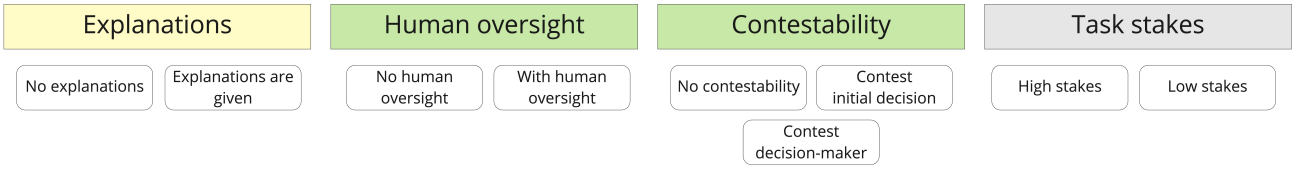

Figure 2: Overview of the independent variables. Yellow refers to information cues, green to system attributes, and grey to contextual factors. White colored boxes indicate the conditions we controlled for each factor.

prediction as an indicator of a potential mistake made by the algorithmic system. The approval process would involve two steps: a first step where the algorithmic system receives an online loan request and evaluates the case; and a second step where a human expert [74] (bank employee) oversees the decision if the algorithmic decision-making system's confidence is low.

- *Contestability* (categorical, between-subjects) We designed contestation mechanisms in the form of appeal processes, following findings from our preliminary study (Section 4.1) and previous literature [68, 101]. Users in our preliminary study mainly wanted to contest (1) the algorithmic decision-maker or (2) the basis of the decision. These strategies resonated with the *new information condition* and *new decision condition* (with a human reviewer) defined by Lyons et al. [68]. We randomly assigned each participant to one of three configurations:

(1) No contestability: participants were told that, due to time constraints, there would be no option for the fictional loan requester to contest the decision in case of a rejection.
(2) Option to contest the initial decision and provide additional information: participants were told that, in case of a rejection, the fictional loan requester had the option to make objections about the initial decision and provide any information to support the application. The same system (if a human oversaw the initial decision, the same human would oversee the review process) would reevaluate the loan application.
(3) Contest decision-maker: participants were told that, in case of a rejection, the fictional loan requester had the opportunity to ask a human (different from the one who oversaw the process if there was already a human involved in the initial decision) to review the process. This human reviewer would make a completely new decision with the information that Kim had already provided for the initial decision.

- *Task stakes* (categorical, between-subjects). Each participant was randomly assigned to one of two configurations:

(1) High-stakes decision: the purpose of the loan application is to buy a house.
(2) Low-stakes decision: the purpose of the loan application is to go on a holiday trip.

4.2.2 Dependent Variables. The instruments we used to measure the dependent variables can be found in our repository.

- *Perceptions of informational fairness* (continuous). Measured by the average score on four of the items used by Schoeffer et al. [83], based on Bies and Moag [15] and Greenberg [40].
- *Perceptions of procedural fairness* (continuous). Measured by the average score on the seven items defined by Colquitt [24],[10] based on Thibaut and Walker [89] and Leventhal [62].
- *Perceptions of overall fairness* (continuous). Measured by a single item rated on a seven-point Likert scale [56, 58].

4.2.3 Descriptive and Exploratory Measurements. The instruments we used to measure the descriptive and exploratory variables can be found in our repository.

- *Age group* (categorical). Participants selected their age group from multiple choices.
- *Level of education* (categorical). Participants selected their highest completed level of education from multiple choices.
- *AI literacy* (continuous). AI literacy has been proven to significantly affect perceptions of informational fairness [83]. We, therefore, captured the average score of the four items defined by Schoeffer et al. [83].

[10] After pilot testing the wording and layout of the presented scenarios, we rephrased some of the items to make them more understandable for participants.

- *Affinity to technology* (continuous). Langer et al. [56] showed that affinity to technology was consistently correlated with end users' perceptions of algorithmic capabilities. We, therefore, captured the average score of the four items defined by Franke et al. [35] as a possible control variable.
- *Personal experience* (continuous). Kramer et al. [55] showed that preferences towards humans vs. algorithmic systems depend on people's previous experience with the described situation. We, therefore, captured the average score of the two items defined by Kramer et al. [55].
- *Task stakes perception* (continuous). Since the stakes involved in a decision are subjective and personal [49], we captured participants' task stakes perceptions as a manipulation check. This was measured through an adapted version of the item defined by Lyons et al. [68].

4.2.4 Procedure. The study consisted of four main steps.

Step 1. Participants stated their age group and level of education. Their degrees of AI literacy, affinity to technology, personal experience and task stakes perception were also measured.

Step 2. Participants were presented with a fictional loan approval scenario involving a person named Kim. Previous research has shown that under *repeated interactions* with algorithmic decision-making systems, decision subjects' fairness perceptions are affected by the favorability of the system towards the group that the decision subjects belong to [37]. In order to minimize these effects, we limited our study to a *one-shot interaction* with the system and we did not disclose the demographics of Kim, such as their gender and age. Kim had applied for a loan online and was waiting for the bank to assess their eligibility. Depending on the stakes of the task that participants had been assigned to, the purpose of this loan would be either to buy a house (high stakes) or to go on a holiday trip (low stakes). Participants would be informed about the information Kim had provided to the bank to evaluate the loan request. As part of the scenario, every participant would then be informed that Kim's loan request had been rejected and they would get to know the process through which the loan request had been evaluated. Based on which of the $(2 \times 2 \times 3 \times 2) = 24$ between-subject scenarios a participant had randomly been placed in, participants would receive explanations about the outcome of the decision, learn whether there was a human expert overseeing the process and get information about whether and how Kim could contest the decision (see Table 2). Participants would then respond to an attention check, where they would be asked about the purpose of the loan request.

Step 3. Participants evaluated their perceptions of informational, procedural, and overall fairness. Additionally, this step included a second attention check that asked participants to select a specific option from a Likert scale.

Step 4. Participants were asked two optional open-ended questions to describe what kind of information they would have liked to receive (if any) and what element would have made the decision-making process fairer (if any).

4.2.5 Data Collection. We planned to collect data from at least 261 participants. We computed the required sample size using the software *G*Power* [34] for an ANOVA with main effects and interactions; specifying the default effect size of 0.25, a significance threshold of $\alpha = \frac{0.05}{11} = 0.0045$ (i.e., due to testing multiple hypotheses; see Section 4.2.6), a desired power of 0.8, 24 groups, and the respective degrees of freedom for the different hypotheses we aimed to test.

We recruited 279 participants from *Prolific* (https://prolific.co). Each participant was at least 18 years old, had high proficiency in English, and could participate in our study only once. Participants were rewarded based on a $12 hourly rate and the median completion time was 7 minutes and 41 seconds. Participants were excluded from data analysis if they did not pass at least one of the attention checks in the experiment. This led to a total number of 267 participants. The study itself was conducted on *Qualtrics* (https://www.qualtrics.com), where participants authenticated with a registration token received on *Prolific*. Our study was approved by a research ethics committee at our institution.

4.2.6 Statistical Analyses. Before conducting any statistical analyses, we mapped all (seven-point) Likert scale answers onto an ordinal scale ranging from −3 (i.e., strongly disagree) to 3 (i.e., strongly agree) and computed averages for answers on related items (e.g., to obtain participants' informational and procedural fairness perceptions).

We analyzed the hypotheses we specified in Section 3 in three separate statistical analyses. First, to test $\mathbf{H_{1a}}$ and $\mathbf{H_{2a}}$, we conducted a multi-way ANOVA with *explanations*, *human oversight*, *contestability*, and *task stakes* as between-subjects factors and *perceptions of informational fairness* as dependent variable.[11] Second, to test $\mathbf{H_{1b\text{-}e}}$ and $\mathbf{H_{2b\text{-}c}}$, we conducted another multi-way ANOVA with the same between-subjects factors but with *perceptions of procedural fairness* as the dependent variable. Third, to test $\mathbf{H_{3a\text{-}c}}$, we conducted a multiple linear regression analysis with *perceptions of informational fairness* and *perceptions of procedural fairness* as independent and *perceptions of overall fairness* as dependent variables. Because we were testing 11 hypotheses as part of this study, we applied a Bonferroni correction to our significance threshold, reducing it to $\frac{0.05}{11} = 0.0045$. This means that p-values resulting from the analyses described above are only regarded as significant if they are below this reduced threshold. Next to the F statistic and p-value, we also report the partial eta squared (η_p^2) effect size for each hypothesis test that was part of an ANOVA.

In addition to the analyses described above, we conducted posthoc tests (i.e., to analyze pairwise differences), Bayesian hypothesis tests[12] (i.e., to quantify evidence in favor of null hypotheses), and exploratory analyses (i.e., to note any unforeseen trends in the data) to better understand our results. We also performed a qualitative, reflexive thematic analysis [19]. The first author coded the responses to the open-ended questions inductively using *Atlas.ti* (https://atlasti.com). These codings were grouped into themes and iteratively refined.

[11] Although we did not specifically hypothesize about the effects of human oversight and contestability on informational fairness perception, we included these variables here for exploratory analyses.

[12] Depending on the outcome of the relevant classical hypothesis test, we report Bayes Factors in favor of the alternative hypothesis (BF_{10}) or the null hypothesis (BF_{01}). We interpret the Bayes Factors according to the guide by Lee and Wagenmakers [57] who adapted it from Jeffreys [46].

A bank has implemented a new loan application system where potential customers apply for a loan online and then the company assesses the eligibility of the customer for the loan.

<Configuration [*No human oversight*] or [*With human oversight*]>

Kim, a potential customer, is looking for funding opportunities to <task> and has thus decided to apply for a <task> loan through the bank's online platform. As part of the <task> loan application process, the bank has requested the following information:

- Applicant annual income
- Co-applicant (if any) annual income
- Credit score
- Date of birth
- Employment status
- Education
- Loan amount requested
- Loan amount term (months)
- Loan purpose
- Number of dependents

A few hours after sending the requested information, Kim has received an email with the final decision: the loan has been rejected.

<Configuration [*No explanation*] or [*With explanations*]>

<Configuration [*No contestability*] or [*Contest initial decision*] or [*Contest decision-maker*]>

Table 2: Overview of the scenario.

5 RESULTS

In this section, we analyze the results of the main user study (see Section 4.2). Table 3 shows a summary of our results.

5.1 Descriptive Statistics

Of the 267 participants in our user study, 19.5% were between 18 and 25 years old, 35% between 26 and 35 years old, 28.5% between 36 and 50 years old, and 17% were between 50-80. 60% of the participants had at least a Bachelor's degree. 87% of our participants claimed to have heard or had experience with humans making loan decisions, whereas 72% of them had heard of or had experience with an algorithmic system making the decision.

5.2 Hypothesis Tests

Our first confirmatory analysis was a multi-way ANOVA with the presence of explanations, human oversight, contestability, and task stakes as between-subjects factors and perceptions of informational fairness as the dependent variable. We found a main effect of the presence of *explanations* ($\mathbf{H}_{1a}$; $F(1, 260) = 74.21$, $p < 0.001$, $\eta_p^2 = 0.22$; $BF_{10} > 1000$) on end users' informational fairness perceptions. However, we did not find any evidence indicating that the effect of explanations on informational fairness is moderated by *task stakes* ($\mathbf{H}_{2a}$; $F(1, 260) = 0.01$, $p = 0.92$, $\eta_p^2 < 0.01$). A Bayesian analysis revealed moderate evidence in favor of the null hypothesis that there is no such interaction effect ($BF_{01} = 7.44$).

The second multi-way ANOVA analysis we conducted had the presence of explanations, human oversight, contestability, and task stakes as between-subjects factors and perceptions of procedural fairness as the dependent variable. We did not find any evidence of *human oversight* impacting procedural fairness perceptions ($\mathbf{H}_{1b}$; $F(1, 254) = 0.004$, $p = 0.95$, $\eta_p^2 < 0.01$) and a Bayesian analysis returned moderate evidence in favor of the null hypothesis that human oversight has no effect here ($BF_{01} = 7.43$). However, there was a strong effect of *contestability* ($\mathbf{H}_{1c}$; $F(2, 254) = 20.60$, $p < 0.001$, $\eta_p^2 = 0.14$; $BF_{10} > 1000$). We further found no evidence in favor of the effect of *contestability* on end users' perceptions of procedural fairness being moderated by the presence of *explanations* ($\mathbf{H}_{1d}$; $F(2, 254) = 0.16$, $p = 0.85$; $\eta_p^2 < 0.01$, $BF_{01} = 12.95$) or by the presence of *human oversight* ($\mathbf{H}_{1e}$; $F(2, 254) = 0.005$, $p = 1.00$; $\eta_p^2 < 0.01$, $BF_{01} = 13.35$). We also did not find any evidence of an interaction between *task stakes* and *human oversight* ($\mathbf{H}_{2b}$; $F(1, 254) = 0.06$, $p = 0.80$, $\eta_p^2 < 0.01$; $BF_{01} = 7.32$) or *task stakes* and *contestability* ($\mathbf{H}_{2c}$; $F(2, 254) = 0.52$, $p = 0.60$, $\eta_p^2 < 0.01$; $BF_{01} = 7.20$) when predicting perceptions of procedural fairness.

We performed a multiple linear regression analysis to test the association of informational and procedural fairness perceptions with overall fairness perceptions ($R^2 = 0.46$, $F(3, 263) = 76.02$, $p < 0.001$). Our results show that *perceptions of informational fairness* ($\mathbf{H}_{3a}$; $\beta = 0.27$, $p < 0.001$) and *perceptions of procedural fairness* ($\mathbf{H}_{3b}$; $\beta = 0.87$, $p < 0.001$) both predicted overall fairness perceptions, with procedural fairness perceptions being the stronger predictor. However, we did not find evidence that perceptions of informational and procedural fairness interact ($\mathbf{H}_{3c}$; $\beta = -0.09$, $p = 0.07$) when predicting overall fairness perceptions.

In sum, we found evidence in favor of four of our hypotheses: $\mathbf{H}_{1a}$, $\mathbf{H}_{1c}$, $\mathbf{H}_{3a}$, and $\mathbf{H}_{3b}$, indicating effects of explanations on informational fairness perceptions and contestability on procedural fairness perceptions, respectively (Figure 3). We also show that informational and procedural fairness perceptions are positively related to overall fairness perceptions.

5.3 Exploratory Analyses

In addition to the hypothesis tests (see Section 5.2), we performed several exploratory analyses to better understand our results and identify any unforeseen but interesting trends in our data. Note that these are not confirmatory results as we did not preregister any of the analyses presented in this subsection.

Decision tasks are subjective and personal [49], so we conducted a manipulation check regarding the stakes of the task. We performed a *t*-test between the pre-defined task stakes (low for a holiday loan, high for a home loan) and participants' perceived task stakes. Our results indicate that the holiday loan ($M = 0.38$, $SD =$

	Informational Fairness					Procedural Fairness								Overall Fairness
	Mean	Th	R	T	U	Mean	V	Inf	Cnst	LB	AF	Crr	Eth	
Explanations	***	◇	◇	◇	◇	◇		◇	◇				◇	
Explanations × Task Stakes														
Explanations × AI literacy	◇		◇											
AI literacy	◇				◇									
Human Oversight														
Human Oversight × Task Stakes														
Contestability						***	◇				◇	◇		
Contestability × Explanations									◇					
Contestability × Human Oversight		◇					◇	◇						
Contestability × Task Stakes											◇			
Task Stakes											◇			
Informational Fairness Perceptions														***
Procedural Fairness Perceptions														***

Table 3: Summary of our results. * refer to confirmatory results ($p < 0.001$), whereas ◇ refer to exploratory results ($p < 0.05$). Empty cells indicate an absence of significant effect between variables. Mean = averaged value of the sub-items that constitute faceted fairness perceptions, Th = Thorough, R = Reliable, T = Tailored, U = Understandable, V = procedural Voice, Inf = Outcome Influence, Cnst = process Consistency, LB = Lack of Bias, AF = Adequacy of Factors, Crr = Correctability, Eth = Ethicality.**

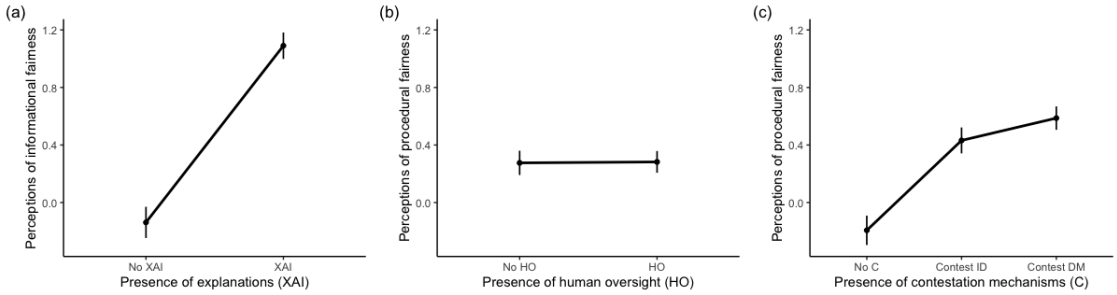

Figure 3: Effects of (a) explanations on perceptions of informational fairness and, (b) human oversight, and (c) contestability on perceptions of procedural fairness (HO = human oversight, C = contestability, ID = initial decision, DM = decision-maker).

1.31) was, indeed, regarded as a lower-stakes scenario compared to the home loan ($M = 1.70$, $SD = 1.07$; $t(258.61) = 9.09, p < 0.001$).

Because contestability is composed of three different groups, we performed pairwise comparisons to analyze the specific differences with respect to procedural fairness perceptions. We observed no significant difference between the effect that the two suggested contestation mechanisms have on procedural fairness perceptions (Tukey-adjusted $p = 0.45$), but both of them differed from the option with no contestability (Tukey-adjusted $p < 0.001$ in both cases).

We also looked at the effects of explanations, human oversight, and contestability on the sub-elements of informational and procedural fairness perceptions. Each of these sub-elements is assessed by one individual item in the fairness perception questionnaires. For *informational fairness perceptions*, we evaluated whether participants thought that Kim received (1) thorough, (2) reasonable, (3) tailored, and (4) understandable information. For *procedural fairness perceptions* we evaluated perceptions of (1) procedural voice, (2) influence over the outcome, (3) consistency of the process, (4) lack of bias, (5) accuracy of factors, (6) correctability, and (7) ethicality. We thus performed multi-way ANOVAs with explanations, human oversight, contestability, and task stakes as between-subjects factors, and the sub-elements that compose informational and procedural fairness perceptions as the dependent variables.

5.3.1 Effects of Explanations. As expected, providing explanations had a positive effect on end users' perceptions of informational fairness. Participants considered that, whenever explanations were added, the bank was giving thorough ($F(1, 249) = 104.00, p < 0.001$, $\eta_p^2 = 0.29$) and reasonable ($F(1, 249) = 40.31, p < 0.001, \eta_p^2 = 0.14$) information that would make Kim understand ($F(1, 249) = 19.84, p < 0.001$, $\eta_p^2 = 0.07$) the way in which the decision was made. Participants also considered that these explanations were tailored to Kim's needs ($F(1, 249) = 45.55, p < 0.001$, $\eta_p^2 = 0.15$). The effect on procedural fairness was partial: our exploratory analysis suggests that explanations affected perceptions of process consistency ($F(1, 254) = 16.80, p < 0.001$, $\eta_p^2 = 0.06$), potentially because explaining to end users how each factor contributes to a final decision may make them discover that the process is standardized and uses the same criteria for every client. Explanations also seemed to interact with contestability in perceptions of procedural consistency ($F(2, 254) = 3.83, p < 0.05$, $\eta_p^2 = 0.03$). Moreover, we checked the interaction of AI literacy and explanations on informational fairness perceptions by performing a multi-way ANOVA with explanations, human oversight, contestability, task stakes, and AI literacy as between-subject factors and perceived informational fairness as the dependent variable. We found that *AI literacy* may have an effect on perceptions of informational fairness

($F(1, 249) = 4.14, p < 0.05, \eta_p^2 = 0.02$) and that *explanations* and *AI literacy* may interact ($F(1, 249) = 4.19, p < 0.05, \eta_p^2 = 0.02$) in creating perceptions of informational fairness (see Figure 5). These results suggest that participants with low AI literacy rated informational fairness perceptions negatively when no explanations were given, but their perceptions of informational fairness substantially increased when decisions were explained. The presence of explanations had a milder effect on informational fairness perceptions among participants with higher AI literacy.

5.3.2 Effects of Human Oversight. Our exploratory analyses suggest that human oversight had no effect on any of the items that contribute to procedural fairness perceptions individually. As a matter of fact, our results show that the inclusion of human oversight in the initial decision has a slight negative impact on perceptions towards process consistency and lack of bias (Figure 4). Human oversight and contestability further seemed to interact in affecting procedural voice perceptions ($F(2, 254) = 4.08, p < 0.05, \eta_p^2 = 0.03$) and outcome influence ($F(2, 254) = 3.65, p < 0.05, \eta_p^2 = 0.03$). This result may suggest that configurations where decision subjects can contest the decision basis of the process lead to varying degrees of procedural voice and outcome influence perceptions depending on whether the initial decision was overseen by a human or not.

5.3.3 Effects of Contestability. In our exploratory analysis, we found that contestability mainly contributed to the "correctability" sub-element of procedural fairness perceptions ($F(2, 254) = 108.29, p < 0.001, \eta_p^2 = 0.46$). This is somewhat unsurprising considering that correctability directly refers to the requirement of having an appeal process in place [62]. Interestingly, however, although contestability seemed to improve perceptions of procedural voice ($F(2, 254) = 13.76, p < 0.001, \eta_p^2 = 0.1$), the mean values of perceived procedural voice are still below zero (on a $[-3, 3]$ scale) for all three configurations: the configuration where there is no contestability ($M = -1.84, SD = 0.16$), the configuration where participants can contest the initial decision ($M = -0.81, SD = 0.17$) and the configuration where participants can contest the decision-maker ($M = -0.65, SD = 0.19$) (Figure 4). The mean values for perceptions of outcome influence are also below zero for all three configurations: no contestability ($M = -1.69, SD = 0.16$), contest initial decision ($M = -1.21, SD = 0.16$) and contest decision-maker ($M = -1.30, SD = 0.16$). This suggests that none of the contestation mechanisms put in place may sufficiently contribute to users' sense of having a voice in the process and influence over the outcome (i.e., the first two sub-elements that constitute procedural fairness perceptions). Our exploratory results also do not point to any differences between contestation types for any of the sub-elements that compose procedural fairness perceptions; except for ethicality ($\beta = -0.81, p < 0.05$). This might indicate that, based on ethical and moral standards, participants do require human intervention in the review process. Note that there is no interaction between contestation types and human oversight for ethicality, which could suggest that having a human-in-the-loop configuration in the original decision is no substitute for human intervention in the review process when upholding ethical standards.

5.3.4 Effects of Task Stakes. Our exploratory analyses surprisingly suggest that task stakes contribute to one item of procedural fairness perceptions: adequacy of factors (e.g., credit score, loan amount requested, total annual income) ($F(1, 254) = 86.79, p < 0.001, \eta_p^2 = 0.25$; see Figure 5). This suggests that users perceived the decision factors used in our scenario as less adequate for the low-stakes decision (holiday) than for the high-stakes decision (buying a house).

5.4 Qualitative Analysis

We performed our qualitative analysis using a reflexive thematic analysis [19]. We inductively generated individual codes from the responses our participants gave to the open-ended questions and we then clustered them into **code groups**. We identified three main tension areas: one related to perceptions of informational fairness and two related to perceptions of procedural fairness. This section explains each of those areas of tension in detail. For a comparison and discussion between quantitative and qualitative results, see Sections 6.1, 6.2, and 6.3. We again refer to quotes as Q.*i*, where *i* is the index of a specific quote. Appendix A shows all selected quotes.

5.4.1 Tension #1: Amount of Information vs. Generating Understanding for All. Our qualitative results indicate that getting detailed information about the decision was a general concern among participants. Participants who were placed in a configuration without explanation of the decision outcome directly highlighted the need for the bank to give **detailed explanations** (115/133) about the way in which different factors are used for making the decision and the reasons for the outcome (Q.11). They also considered that the bank should provide decision subjects with an alternative **course of action** (34/133; Q.12).

Participants who were placed in scenarios where the bank would offer explanations of the decision outcome positively evaluated the level of detail of this information (70/134). They generally also appreciated the fact that the counterfactual scenarios gave actionable information (21/134). Some of them requested **further information about the process and the algorithmic system** itself (51/134; Q.13). However, some participants pointed out that increasing the amount of information could generate **difficulties in understanding** (23/134) the explanations and could restrict such understanding to people with literacy in AI (Q.14).

5.4.2 Tension #2: Human Involvement vs. Timely Decision-Making. Another major theme in our qualitative analysis was that of human involvement. Our qualitative analysis suggests that, regardless of the presence or absence of human oversight, participants were still asking for a **higher degree of human involvement** (75/267) in the process (e.g., by including a human that deals with borderline cases, or by allowing decision subjects to personally interact with a bank employee). In cases where human oversight was included in the original decision, our participants thought that this would ensure reliability. However, some (13/267) of them indicated that a human should always make the final decision, for every instance (Q.15, Q.16).

On the other hand, as some of our participants highlighted, not having humans involved could make the **process speedy** (47/267) and would allow Kim to explore alternative options (Q.17). Although we did not explicitly compare the difference in time of having a

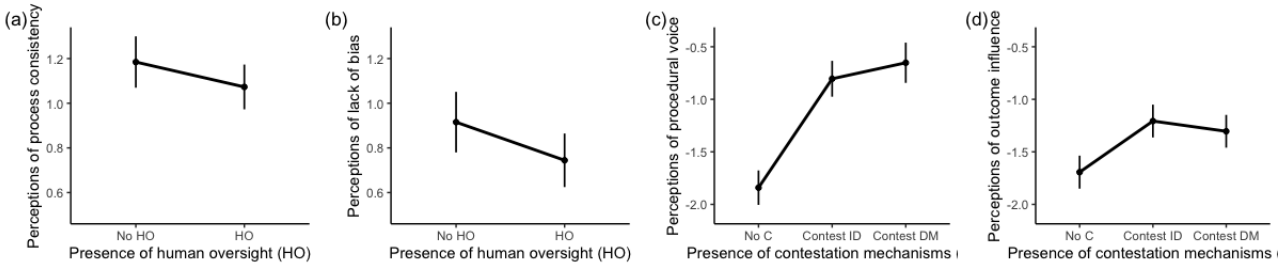

Figure 4: Effects of human oversight on perceptions of (a) process consistency and (b) lack of bias; effects of contestability on perceptions of (c) procedural voice and (d) outcome influence (HO = human oversight, C = contestation, ID = initial decision, DM = decision-maker).

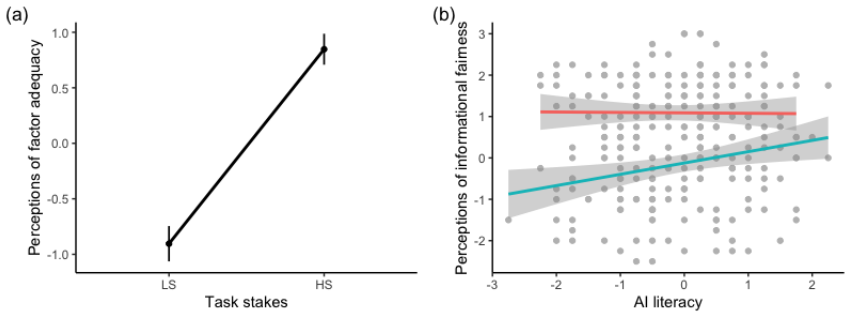

Figure 5: (a) Effect of task stakes on perceptions of factor adequacy (LS = Low stakes, HS = High stakes). (b) Interaction between explanations and self-reported AI literacy on perceptions of informational fairness. Red refers to the configurations where explanations were given and Green refers to the configurations with no explanations.

human or an algorithmic system (with or without human oversight) making the decision, the presented scenario did mention that the reason for introducing algorithmic decision-making processes was due to time constraints. Many participants referred to the temporal dimension as one that makes the process fair (Q.18, Q.19).

5.4.3 Tension #3: Standardized Fact-based Process vs. Accounting for Personal Circumstances. The fact that an algorithmic system was fully or mainly driving the process also encouraged reflections on the advantages and disadvantages of having a standardized process that treats **everyone equally** (44/267; Q.20). Some of our participants considered that introducing algorithmic systems in decision-making processes helps to **get rid of human biases** (39/267). They considered that thanks to such systems, the process would not be subject to human subjectiveness and prejudice (Q.21). Introducing an algorithmic system was also viewed as contributing to the consistency of the decision-making process. Participants generally appreciated that the same information was considered for everyone (Q.22). The basis of the decision-making process was also regarded as sound because it was **based on facts** (40/267; Q.23). Some (27/267) indicated that the bank should consider additional factors when making a decision, but, in general terms, the presented factors were considered fair and relevant (Q.24).

Despite the general sentiment of facts being a sound basis for decision-making, some of our participants highlighted the need to sometimes consider **individual circumstances** (17/267; Q.25, Q.26). Humans were viewed as being more flexible and prone to give in to cases that are close to the decision boundary (Q.27). Some participants pointed out that a human should be responsible for double-checking boundary cases (Q.28). In those cases, participants requested the implementation of negotiation mechanisms (Q.29) that would allow decision subjects to **discuss with humans** (47/267; Q.30) who could treat the situation with compassion (Q.31).

6 DISCUSSION

In this section, we relate quantitative results with qualitative ones and reflect on our key findings. Each subsection summarizes the results related to one of the tested factors and its entanglements (i.e., explanations in Section 6.1, human oversight in Section 6.2, and contestability in Section 6.3). We also list the practical implications of our findings, highlight future challenges, and reflect on the benefits and shortcomings of adopting a multi-dimensional approach for capturing perceptions of fairness (Section 6.4). We finally acknowledge the limitations of our study (Section 6.5).

6.1 Leveraging Transparency Beyond Outcome Explanations

Our quantitative results show that explanations improve informational fairness perceptions (see Section 5.2). Exploratory findings

further suggest that AI literacy may moderate the effect of explanations on informational fairness perceptions, i.e., indicating that the effect of explanations on informational fairness perceptions is stronger for participants with low AI literacy (see Section 5.3.1 and Figure 5). However, contrary to our expectations, and to suggestions from earlier work [83], we did not find evidence that explanations moderate the effect of contestability on procedural fairness, i.e., help participants question structural aspects of the decision-making process such as the factors requested by the bank and how these are used. The insights we obtained from our qualitative analysis suggest that participants were generally happy with the factual basis of the decision in question (see Section 5.4). It should be noted that, as opposed to earlier work [83] and our own preliminary study, we had decided to discard gender as one of the decision-making factors in our main study because it is explicitly protected by law [14]. This might have influenced how people perceived the decision basis. Moreover, some participants were asking for system-level explanations that would enable them to explore and evaluate biases encoded in the algorithmic system. The lack of this information might have prevented them from questioning additional aspects of the decision-making.

Implications. Although our study replicated the finding from earlier work that explanations support informational fairness perceptions [83] (which in turn contribute to overall fairness perceptions), restricting explanations to technical solutions that are currently available through XAI may limit the grounds for contestations [67]. Our results (e.g., Q.13) suggest that providing decision subjects with information that goes beyond outcome explanations could support contestations that are not only limited to post-decision mechanisms but that apply to the system lifecycle as a whole [2]. These system-level explanations could include information about data, algorithmic features, or the way in which algorithmic systems are integrated in broader workflows [30]. For instance, previous studies have shown that data-centric explanations [7] have the potential to assist users in assessing fairness. Future work should look into explanations and transparency that go beyond outcomes and test how these insights affect perceptions of informational fairness and whether they set grounds for contestations that go beyond appeal processes. We foresee that this would not only have implications for perceptions of informational fairness but also for perceptions of procedural fairness.

Challenges. Previous research has demonstrated that increasing levels of transparency can lead to information overload [22], so expanding explanations could restrict understanding to individuals with literacy in AI. Moreover, earlier work has pointed to a risk that malicious actors might use explanations to defraud algorithmic systems [105] or to manipulate decision subjects by conveying untruthful levels of "fairness" [68]. Future work should look into methods for designing strategies that leverage adequate levels of transparency [105] and that convey *appropriate fairness perceptions* (i.e., condition that is satisfied if fairness perceptions towards a system are high when the system is indeed fair) [82]. Such strategies should be adapted to decision subjects' insight needs [88] and designed in a way that they would understand [11, 52]. For example, these could include videos [93], stories [93], or comics [102, 104]. Our qualitative analysis further revealed some participants' feeling that the process could not be biased because it is impossible for algorithmic systems to be biased (Q.20), suggesting that future explanations should also account for decision subjects' imaginaries [69] and expectations [54] around algorithmic systems.

6.2 Designing Appropriate Human-AI Configurations

Our quantitative results do not contain any evidence that human oversight would affect end users' procedural fairness perceptions; in fact, a Bayesian analysis even revealed moderate evidence that human oversight has no effect here (see Section 5). These results resonate with earlier work on the topic [103], where a case-by-case human intervention did not contribute to perceptions of fairness. Nevertheless, our qualitative results suggest that, regardless of human oversight in the original decision, participants were still asking for a higher degree of human intervention (e.g., Q.15; see Section 5.4). The reason for this might be that end users might think about the decision-maker in binary terms, as either "a human" or "not a human" [56]. Since, even in the scenario with human oversight, the first prediction was made by the algorithmic system, our participants might still have thought about it as a non-human decision-maker. This would explain why human oversight did not affect perceptions of procedural fairness and why, even in the case where the decision was overseen by a human, participants were asking for more human intervention in the process.

Implications. More research is needed to find adequate forms of human-AI collaborations in algorithmic decision-making processes. Future studies should go beyond configurations where humans confirm the quality of the decision made by an algorithmic system [5] and craft alternative human-AI teams. For instance, algorithmic systems could access large quantities of data and perform preliminary analyses to produce easily digestible summaries for human experts to make final decisions [76]. Such a configuration would respond to our participants' desire to always have a human making the last decision. A follow-up study to ours could test perceptions towards human decision-making processes that are advised by algorithmic systems [12, 109] rather than algorithmic decision-making processes that are overseen by humans. One could argue that many studies have already studied different human-AI teaming configurations. However, these studies have mainly focused on exploring the interaction of data domain experts (i.e., bank employees in our case) with algorithmic systems and distilling the effect on trust [75, 81] or trust-related constructs [100] such as reliance [77, 109]. Future studies should also capture end users' fairness perceptions for each of those configurations.

Challenges. Including humans in algorithmic decision-making processes costs time [20, 29, 68] and our qualitative results suggest that participants value timely decision-making processes. For appeal processes, Lyons et al. [68] found that, when subject to a trade-off situation, participants prioritised the type of review and the review time rather than the reviewer. We emphasize the need to perform more studies where participants are shown the time cost of different configurations so as to capture their perceptions of procedural fairness in a space of trade-offs. Furthermore, our participants regarded configurations with no human intervention as less biased and more consistent. We echo Almada [5] and suggest that comparative measures of performance of human-controlled and

fully automated procedures should be included. This would allow end users to freely shape their preferences and fairness perceptions in an informed way.

6.3 Giving Voice to Decision Subjects

As we hypothesized, our quantitative results show that including contestability (in the form of appeal processes) enhances people's perceptions of procedural fairness. Our qualitative results back up the value that participants put on the ability to contest the decision. Despite the positive effect of contestability on perceptions of procedural fairness, perceptions of procedural voice and influence over the outcome were still negative. In a within-subjects user study, Lyons et al. [68] found that participants perceived the *new information* appeal condition (equivalent to our "option to contest the initial decision and provide additional information" appeal condition) as fairer than the rest of the suggested appeal processes. Contrary to these findings, we do not find any differences between the suggested appeal processes. This might be due to the between-subject nature of our study. Lyons et al. [68] also found that the reason for the preference towards this condition was that decision subjects perceived they had a "voice" in the decision-making process. Our results contradict these findings, and indicate that, even when any of the suggested appeal processes are in place, our participants did not have the feeling that the decision subject had a voice in the process or influence over the outcome. This discrepancy might be due to the nature of the performed analysis. Lyons et al. [68] arrived at this conclusion through a thematic analysis of qualitative data, whereas our results rely on quantitatively evaluating responses to statements that directly address perceptions of procedural voice and influence over the outcome.

Implications. Our findings highlight that, although contestability enhances users' perceptions of procedural fairness (which in turn contribute to overall fairness perceptions), more research in contestable AI is needed. The field of contestable AI is still growing [3] and many of the guidelines on how to design for contestability are conceptual in nature [3, 44, 67]. Further research is necessary to translate those conceptualizations into actual design guidelines [2, 60] and validate designs of contestable algorithmic systems. Our results also suggest the need to research into the design of contestation mechanisms that effectively provide voice and outcome influence to decision subjects. Sarra [79] argue that a "dialectical exchange" is necessary between decision subjects and human controllers to effectively support contestability. This resonates with our qualitative findings: many of our participants were asking for options to personally discuss or negotiate the outcomes with humans. Our participants considered that discussing the decision with humans would potentially lead to a change in outcome for cases that were close to the decision boundary (e.g., Q.27, Q.28; in line with earlier work [36, 68]) and that humans would treat decision subjects with dignity and compassion (e.g., Q.31; also in line with previous research [17, 68, 94]). These findings further suggest that contestations might be better designed as dialogues [44, 53], rather than mere appeal processes. When it comes to outcome influence, future research should focus on ways of increasing the ability of subjects to exercise agency and true influence over the process [9]. This entails allowing decision subjects to determine the input data that they want to provide along with the ability to influence the logics of the decision-making process [60]. A promising research line in this field is that of *interactive contestations* [45].

Challenges. A major challenge when trying to give effective outcome influence to decision subjects is the distribution of levels of control across individuals. Since the process will eventually influence multiple people rather than one individual, the way in which this control is distributed remains a key challenge [71]. We consider that participatory design strategies [43], such as the workshops conducted by Vaccaro et al. [94], can help deal with the trade-offs identified in our qualitative analysis. These workshops facilitate conversations among different stakeholders (e.g., the development team and decision subjects) and could, therefore, help identify the compromises in designing contestation mechanisms that attend to individual circumstances while contributing to perceptions of process consistency.

6.4 Multi-dimensional Measurement of Fairness Perceptions

In this paper, we advocated for a multi-dimensional approach for capturing perceptions of fairness, inspired by literature in human decision-making. Our quantitative analyses confirm that informational and procedural facets of fairness predict overall fairness perceptions. Moreover, this multi-dimensional approach has enabled us to perform exploratory analyses that have generated a nuanced understanding of how people perceive each algorithmic configuration. Our findings, therefore, suggest that future studies and practical applications could benefit from adopting a multi-dimensional rather than a one-dimensional approach.

Despite our promising findings, using a tool that was designed for human decision-making to evaluate algorithmic decision-making may not encompass the unique challenges that the inclusion of algorithmic systems bring to existing processes (as it is the case for other fields such as human-agent collaboration [23]). Our aim behind using this tool designed for human decision-making in an algorithmic context was to distil insights from it and to identify future research directions. There is evidence that suggests that decision subjects care about justice-related aspects in algorithmic decision-making, as they care in human decision-making [17]. However, we acknowledge that there are novel considerations that the usage of these systems results in [17] and that future work should consider. For instance, the approach suggested by Colquitt [24] does not explicitly include the temporal dimension of the decision-making process as an attribute that contributes to perceptions of procedural fairness. Through our qualitative analysis, we found that this aspect was paramount for our participants. We note that most of the criteria we evaluated were defined several decades ago. Due to societal changes and a change in perceptions of time brought in by algorithmic systems, further research would be needed to consider and effectively evaluate speed of decision-making as a procedural justice principle [95]. We, therefore, encourage further research into defining standardized methodological approaches that appropriately capture perceptions of fairness across dimensions while being specifically adapted to algorithmic decision-making.

6.5 Limitations

In this section, we summarize limitations of our study that could represent threats to its validity.

Reflections on our experimental setting. The design of our study might have had an impact on the obtained results. First, the between-subjects nature of the study might have prevented participants from comparing different algorithmic configurations. The effects of task stakes and human oversight might have been diluted because of this. Second, the scenario used for conducting our controlled user study presented a case that participants considered to be close to the decision boundary (see Q.27). This made the request to have a human involved in the decision-making process, for example, to be especially relevant for some participants (see Q.28). Fairness perceptions and the desires expressed by participants might have been different if we had included scenarios with different characteristics. Third, the design of our experiment described a loan denial scenario for an individual called Kim. As opposed to some other authors (e.g., [68, 103]) we decided to tell this story in the third person [9, 83, 86] with no reference to the individuals' personal characteristics. The reason behind this design choice was to minimize, as far as possible, the *outcome favourability bias* [103]. In the same line, we limited the interaction between participants and the algorithmic system to a one-shot interaction. Previous research has shown that, under repeated interactions, system favorability towards the group that the decision subject belongs to has an effect on fairness perceptions [37]. Our results indicate that, generally speaking, participants were happy to endorse negative outcomes if explanations and contestation mechanisms were in place. However, outcome favourability bias might have resulted in different reactions had we referred to a case where the participants themselves had been denied a loan or had we disclosed the demographics of different individuals and asked participants to repeatedly interact with the algorithmic system. Fourth, although we varied the level of stakes involved in the task and found that perceptions of informational and procedural fairness are robust across stakes, our study is still limited to a loan decision-making scenario. Results may vary depending on the context. Fifth, terminology has been claimed to affect end users' fairness perceptions [56]. Langer et al. [56] suggest that the usage of multi-item measurement tools softens the impact of terminology, an advice we followed when measuring perceptions of informational and procedural fairness. However, results may have been different had we used terms such as *algorithmic system*, *statistical model*, or *computing system* instead of *artificial intelligence*.

Generalizability across cultures. For our study we recruited participants from the Global North whose first language was English. Previous work has shown that cultural and geographical differences play a key role in perceptions towards algorithmic systems [10, 49, 97]. Thus, we acknowledge that our study is subject to representativeness limitations [61].

Need to incorporate empirical ethics as part of broader design frameworks for algorithmic systems. Empirical studies represent a necessary strategy for testing the practical implications of theoretical claims. However, moving towards algorithmic decision-making processes that enhance decision subjects' feelings of justice requires that empirical studies are part of broader efforts to create methodological tools that consider different stakeholders' (including decision subjects) viewpoints in the design and evaluation processes of algorithmic systems [78, 107].

7 CONCLUSION

This paper presented a preregistered user study investigating how varying levels of *explanations*, *human oversight*, and *contestability* for high- and low-stakes algorithmic loan approval scenarios affect users' informational, procedural, and overall fairness perceptions. We found that explanations and contestability affect perceptions of informational and procedural fairness, respectively. We did not find evidence of the effect of human oversight and task stakes on these measurements. We also found that perceptions of informational and procedural fairness, independently, are positively related to perceptions of overall fairness, but their interaction is not significant. Through exploratory and qualitative analyses, we gave further insights into these relationships. Our exploratory analyses indicated that the suggested contestation mechanisms did not effectively contribute to perceptions of procedural voice and outcome control. Our exploratory analyses also pointed out that the suggested human oversight configuration slightly deteriorated perceptions of procedural consistency and lack of bias. Through a qualitative analysis, we found three main areas of tension that highlight the need to assess algorithmic decision-making processes in a space of trade-offs. Our work, therefore, gives insights into how to design algorithmic decision-making processes that foster feelings of justice and addresses some of the HCI challenges that these systems have brought in.

ACKNOWLEDGMENTS

We thank Himanshu Verma, Alejandra Gomez Ortega, Wo Meijer, Di Yan, and Denis Bulygin for valuable feedback on previous versions of this paper. We also thank the anonymous reviewers for their constructive and thoughtful reviews. We would like to express our gratitude to our colleagues at StudioLab and the DCODE Network for helping us pilot test our study.

This work was partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955990 and the EU Horizon 2020 grant (grant 101016233) via the PERISCOPE (Pan-European Response to the ImpactS of COVID-19 and future Pandemics and Epidemics) project.

REFERENCES

[1] J. Stacy Adams. 1965. Inequity in social exchange. In *Advances in experimental social psychology*. Vol. 2. Academic Elsevier, 267–299.
[2] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. 2022. Tensions in transparent urban AI: designing a smart electric vehicle charge point. *AI & SOCIETY* (3 2022). https://doi.org/10.1007/s00146-022-01436-9
[3] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2022. Contestable AI by Design: Towards a Framework. *Minds and Machines* (8 2022). https://doi.org/10.1007/s11023-022-09611-z
[4] Kars Alfrink, T. Turel, A. I. Keller, N. Doorn, and G. W. Kortuem. 2020. Contestable City Algorithms. International Conference on Machine Learning Workshop.
[5] Marco Almada. 2019. Human intervention in automated decision-making. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. ACM, New York, NY, USA, 2–11. https://doi.org/10.1145/3322640.3326699

[6] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. 2019. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 289–295. https://doi.org/10.1145/3306618.3314243

[7] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. https://doi.org/10.1145/3411764.3445736

[8] Karl Aquino. 1995. Relationships among pay inequity, perceptions of procedural justice, and organizational citizenship. *Employee Responsibilities and Rights Journal* 8, 1 (3 1995), 21–33. https://doi.org/10.1007/BF02621253

[9] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (9 2020), 611–623. https://doi.org/10.1007/s00146-019-00931-w

[10] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (11 2018), 59–64. https://doi.org/10.1038/s41586-018-0637-6

[11] Simone Bae, Reeva Lederman, and Tingru Cui. 2022. Understanding User Perception of Explainable Algorithmic Decision-Making Systems: A Systematic Literature Review. (2022).

[12] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (7 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[13] Julian Barling and Michelle Phillips. 1993. Interactional, Formal, and Distributive Justice in the Workplace: An Exploratory Study. *The Journal of Psychology* 127, 6 (11 1993), 649–656. https://doi.org/10.1080/00223980.1993.9914904

[14] Uladzislau Belavusau and Kristin Henrard. 2019. A Bird's Eye View on EU Anti-Discrimination Law: The Impact of the 2000 Equality Directives. *German Law Journal* 20, 05 (7 2019), 614–636. https://doi.org/10.1017/glj.2019.53

[15] R.J. Bies and J. F. Moag. 1986. Interactional Justice: Communication Criteria of Fairness. . *Research on Negotiations in Organizations* 1 (1986), 43–55.

[16] Robert J. Bies and Debra L. Shapiro. 1987. Interactional fairness judgments: The influence of causal accounts. *Social Justice Research* 1, 2 (6 1987), 199–218. https://doi.org/10.1007/BF01048016

[17] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. (1 2018). https://doi.org/10.1145/3173574.3173951

[18] C. Malik Boykin, Sophia T. Dasch, Vincent Rice Jr., Venkat R. Lakshminarayanan, Taiwo A. Togun, and Sarah M. Brown. 2021. Opportunities for a More Interdisciplinary Approach to Measuring Perceptions of Fairness in Machine Learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, New York, NY, USA, 1–9. https://doi.org/10.1145/3465416.3483302

[19] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (1 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[20] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56, 5 (10 2019), 809–825. https://doi.org/10.1177/0022243719851788

[21] David Chan. 2011. Perceptions of fairness. *Research Collection School of Social Sciences* (2011).

[22] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[23] Nazli Cila. 2022. Designing Human-Agent Collaborations: Commitment, responsiveness, and support. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3491102.3517500

[24] Jason A. Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology* 86, 3 (6 2001), 386–400. https://doi.org/10.1037/0021-9010.86.3.386

[25] Jason A Colquitt and Jessica B Rodell. 2015. Measuring Justice and Fairness. In *The Oxford Handbook of Justice in the Workplace*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199981410.013.0008

[26] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

[27] Russell Cropanzano. 2012. *Justice in the Workplace: From theory To Practice*. Vol. 2.

[28] Morton Deutsch. 1975. Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues* (1975), 137–149.

[29] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. https://doi.org/10.1037/xge0000033

[30] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. (1 2019). https://doi.org/10.1145/3301275.3302310

[31] Tim Draws, Zoltán Szlávik, Benjamin Timmermans, Nava Tintarev, Kush R. Varshney, and Michael Hind. 2021. Disparate Impact Diminishes Consumer Trust Even for Advantaged Users. (1 2021). https://doi.org/10.1007/978-3-030-79460-6{_}11

[32] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[33] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. 2020. FaiRecSys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics* 9, 2 (2020), 197–213. https://doi.org/10.1007/s41060-019-00181-5

[34] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (5 2007), 175–91. https://doi.org/10.3758/bf03193146

[35] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (4 2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150

[36] Elena Fumagalli, Sarah Rezaei, and Anna Salomons. 2022. OK computer: Worker perceptions of algorithmic recruitment. *Research Policy* 51, 2 (3 2022), 104420. https://doi.org/10.1016/j.respol.2021.104420

[37] Meric Altug Gemalmaz and Ming Yin. 2022. Understanding Decision Subjects' Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 295–306. https://doi.org/10.1145/3514094.3534201

[38] Jerald Greenberg. 1987. A Taxonomy of Organizational Justice Theories. *The Academy of Management Review* 12, 1 (1 1987), 9. https://doi.org/10.2307/257990

[39] Jerald Greenberg. 1990. Organizational Justice: Yesterday, Today, and Tomorrow. *Journal of Management* 16, 2 (6 1990), 399–432. https://doi.org/10.1177/014920639001600208

[40] J. Greenberg. 1993. The social side of fairness: Interpersonal and informational classes of organizational justice. *Justice in the workplace: Approaching fairness in human resource management*. (1993), 79–103.

[41] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS SYMPOSIUM ON MACHINE LEARNING AND THE LAW 8*.

[42] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[43] Katrina Heijne and Han van der Meer. 2019. *Road Map for Creative Problem Solving Techniques Organizing and facilitating group sessions*. Boom Uitgevers Amsterdam.

[44] Clément Henin and Daniel Le Métayer. 2021. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY* (7 2021). https://doi.org/10.1007/s00146-021-01251-8

[45] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing Contestability. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, New York, NY, USA. https://doi.org/10.1145/3064663.3064703

[46] Harold Jeffreys. 1939. Theory of Probability. (1939). (1939).

[47] Denise Jepsen and John Rodwell. 2009. A New Dimension of Organizational Justice: Procedural Voice. *Psychological Reports* 105, 2 (10 2009), 411–426. https://doi.org/10.2466/PR0.105.2.411-426

[48] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. (7 2019).

[49] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3491102.3517533

[50] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A survey of algorithmic recourse:contrastive explanations and consequential recommendations. *Comput. Surveys* (4 2022). https://doi.org/10.1145/3527848

[51] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability,*

and Transparency. ACM, New York, NY, USA, 353–362. https://doi.org/10.1145/3442188.3445899

[52] Styliani Kleanthous, Maria Kasinidou, Pınar Barlas, and Jahna Otterbacher. 2022. Perception of fairness in algorithmic decisions: Future developers' perspective. *Patterns* 3, 1 (1 2022), 100380. https://doi.org/10.1016/j.patter.2021.100380

[53] Daniel Kluttz, Nitin Kohli, and Deirdre K. Mulligan. 2018. Contestability and Professionals: From Explanations to Engagement with Algorithmic Systems. *SSRN Electronic Journal* (2018). https://doi.org/10.2139/ssrn.3311894

[54] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[55] Max F. Kramer, Jana Schaich Borg, Vincent Conitzer, and Walter Sinnott-Armstrong. 2018. When Do People Want AI to Make Decisions?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 204–209. https://doi.org/10.1145/3278721.3278752

[56] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–28. https://doi.org/10.1145/3491102.3517527

[57] Michael D. Lee and Eric-Jan Wagenmakers. 2014. *Bayesian Cognitive Modeling*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

[58] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (1 2018). https://doi.org/10.1177/2053951718756684

[59] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1035–1048. https://doi.org/10.1145/2998181.2998230

[60] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–26. https://doi.org/10.1145/3359284

[61] Min Kyung Lee and Katherine Rich. 2021. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445570

[62] Gerald S. Leventhal. 1980. What Should Be Done with Equity Theory? In *Social Exchange*. Springer US, Boston, MA, 27–55. https://doi.org/10.1007/978-1-4613-3087-5{_}2

[63] Q. Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. (4 2022). https://doi.org/10.1145/3531146.3533182

[64] E. Allan Lind and Tom R. Tyler. 1988. *The Social Psychology of Procedural Justice*. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4899-2115-4

[65] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (3 2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[66] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to Medical Artificial Intelligence. *Journal of Consumer Research* 46, 4 (12 2019), 629–650. https://doi.org/10.1093/jcr/ucz013

[67] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. (2 2021). https://doi.org/10.1145/3449180

[68] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3517606

[69] Jakub Mlynar, Farzaneh Bahrami, André Ourednik, Nico Mutzner, Himanshu Verma, and Hamed Alavi. 2022. AI beyond Deus ex Machina – Reimagining Intelligence in Future Cities with Urban Experts. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3491102.3517502

[70] Rosanna Nagtegaal. 2021. The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly* 38, 1 (1 2021), 101536. https://doi.org/10.1016/j.giq.2020.101536

[71] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Towards Involving End-users in Interactive Human-in-the-loop AI Fairness. (4 2022).

[72] Gideon Ogunniye, Benedicte Legastelois, Michael Rovatsos, Liz Dowthwaite, Virginia Portillo, Elvira Perez Vallejos, Jun Zhao, and Marina Jirotka. 2021. Understanding User Perceptions of Trustworthiness in E-Recruitment Systems. *IEEE Internet Computing* 25, 6 (11 2021), 23–32. https://doi.org/10.1109/MIC.2021.3115670

[73] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[74] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (3 2022), 1–31. https://doi.org/10.1145/3512929

[75] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–9. https://doi.org/10.1145/3491102.3502104

[76] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (10 2019), 125–134. https://ojs.aaai.org/index.php/HCOMP/article/view/5281

[77] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–52. https://doi.org/10.1145/3411764.3445315

[78] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

[79] Claudio Sarra. 2020. Put Dialectics into the Machine: Protection against Automatic-decision-making through a Deeper Understanding of Contestability by Design. *Global Jurist* 20, 3 (10 2020). https://doi.org/10.1515/gj-2020-0003

[80] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare?. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 99–106. https://doi.org/10.1145/3306618.3314248

[81] Philipp Schmidt and Felix Biessmann. 2020. Calibrating Human-AI Collaboration: Impact of Risk, Ambiguity and Transparency on Algorithmic Bias. 431–449. https://doi.org/10.1007/978-3-030-57321-8{_}24

[82] Jakob Schoeffer and Niklas Kuehl. 2021. Appropriate Fairness Perceptions? On the Effectiveness of Explanations in Enabling People to Assess the Fairness of Automated Decision Systems. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New York, NY, USA, 153–157. https://doi.org/10.1145/3462204.3481742

[83] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. (5 2022). https://doi.org/10.1145/3531146.3533218

[84] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (11 2017), 233–242. https://doi.org/10.1093/idpl/ipx022

[85] Debra L. Shapiro, E.Holly Buttner, and Bruce Barry. 1994. Explanations: What Factors Enhance Their Perceived Adequacy? *Organizational Behavior and Human Decision Processes* 58, 3 (6 1994), 346–368. https://doi.org/10.1006/obhd.1994.1041

[86] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, New York, NY, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664

[87] Emily Sullivan and Philippe Verreault-Julien. 2022. From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 712–722. https://doi.org/10.1145/3514094.3534185

[88] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445088

[89] J. W. Thibaut and L. Walker. 1975. Procedural Justice: A Psychological Analysis. *L. Erlbaum Associates, Hillsdale*. (1975).

[90] Tom R. Tyler. 1988. What is Procedural Justice?: Criteria used by Citizens to Assess the Fairness of Legal Procedures. *Law & Society Review* 22, 1 (1988), 103. https://doi.org/10.2307/3053563

[91] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 10–19. https://doi.org/10.1145/3287560.3287566

[92] Kristen Vaccaro, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. Contestability in Algorithmic Systems. In *Conference Companion*

Publication of the 2019 on Computer Supported Cooperative Work and Social Computing. ACM, New York, NY, USA, 523–527. https://doi.org/10.1145/3311957.3359435

[93] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants". *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (10 2020), 1–22. https://doi.org/10.1145/3415238

[94] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–28. https://doi.org/10.1145/3476059

[95] Annukka Valkeapää and Tuija Seppälä. 2014. Speed of Decision-Making as a Procedural Justice Principle. *Social Justice Research* 27, 3 (9 2014), 305–321. https://doi.org/10.1007/s11211-014-0214-6

[96] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445365

[97] Niels van Berkel, Eleftherios Papachristos, Anastasia Giachanou, Simo Hosio, and Mikael B. Skov. 2020. A Systematic Assessment of National Artificial Intelligence Policies: Perspectives from the Nordics and Beyond. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3419249.3420106

[98] Arnaud Van Looveren and Janis Klaise. 2021. Interpretable Counterfactual Explanations Guided by Prototypes. 650–665. https://doi.org/10.1007/978-3-030-86520-7{_}40

[99] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 284–293. https://doi.org/10.1145/3351095.3372876

[100] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–39. https://doi.org/10.1145/3476068

[101] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. (11 2017).

[102] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[103] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376813

[104] Zezhong Wang, Jacob Ritchie, Jingtao Zhou, Fanny Chevalier, and Benjamin Bach. 2021. Data Comics for Reporting Controlled User Studies in Human-Computer Interaction. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2 2021), 967–977. https://doi.org/10.1109/TVCG.2020.3030433

[105] Elizabeth Anne Watkins. 2021. The tension between information justice and security: Perceptions of facial recognition targeting.. In *Joint Proceedings of the ACM IUI 2021 Workshops*.

[106] Jenny S. Wesche and Andreas Sonderegger. 2019. When computers take the lead: The automation of leadership. *Computers in Human Behavior* 101 (12 2019), 197–209. https://doi.org/10.1016/j.chb.2019.07.027

[107] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 535–563. https://doi.org/10.1145/3531146.3533118

[108] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 335–340. https://doi.org/10.1145/3278721.3278779

[109] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–28. https://doi.org/10.1145/3491102.3517791

[110] Jianlong Zhou, Sunny Verma, Mudit Mittal, and Fang Chen. 2021. Understanding Relations Between Perception of Fairness and Trust in Algorithmic Decision Making. (9 2021).

A SELECTED QUOTES

Selected quotes from the preliminary study (S1; see Section 4.1) and the main study (S2; see Section 4.2). Each quote comes with a reference to the study where the response was collected and to the the participant (P*j*) who gave it.

Q.id	Quote	Participant
Q.1	*"It is unfair for her to be denied based on someone else's previous inability to pay back the loan"*	S1-P42
Q.2	*"Just because some had a similar case as hers, does not prove that she would not be able to pay back the loan."*	S1-P36
Q.3	*"The best explanation gives the largest volume of information including how the decision was made and what amount she could potentially lend"*	S1-P50
Q.4	*"It explains the importance of each factor so she is able to see clearly what factors are most influential"*	S1-P32
Q.5	*"It boils it down to very easy to digest reasons as to why Kim was rejected the loan request"*	S1-P29
Q.6	*"It provides 3 different ways in which Kim could improve her chances of being accepted."*	S1-P33
Q.7	*"She should contest how little impact her employment has on the decisions since this is a big factor"*	S1-P22
Q.8	*"Gender should be contested as is a discriminatory factor. Although all the variables in question are methods for the banks to discriminate against someone, gender is not within a person's control and therefore a bad measure of their character and choices."*	S1-P56
Q.9	*"Artificial intelligence does not take your lifestyle and circumstances into account."*	S1-P46
Q.10	*"It is assessing her by comparing her situation with another with similar salary & credit score & not taking her full circs [circumstances] into consideration."*	S1-P53
Q.11	*"I think there should be a breakdown of what the artificial intelligence looks for and what the decision is based on."*	S2-P5
Q.12	*"They should offer a detailed reason and list of suggested changes she could make to help her in her efforts"*	S2-P218
Q.13	*"It does not tell us enough about how the AI uses the information. The AI is programmed initially by a human. How can I be sure that no bias is involved in this programming of the algorithm? This would be appropriate information to have."*	S2-P8
Q.14	*"If Kim is not familiar with AI then she may not understand the process and view it negatively"*	S2-P135
Q.15	*"[...] each application should be reviewed by a human, not just the ones which have low confidence"*	S2-P179
Q.16	*"Maybe for it to be processed primarily by the AI but secondly by a human before the answer is finalised. This could still be a quick process as the person wouldn't have to spend much time on it but it would mean the decision also had a human input."*	S2-P226
Q.17	*"It is fairer than other options as [it] is quicker than a human decision - [it] allows customers to explore other options"*	S2-P153
Q.18	*"It is fair because with the help of its AI the application process is much faster and efficient"*	S2-P146
Q.19	*"I do think it is fair, it is a quick and easy procedure"*	S2-P182
Q.20	*"It's fair because it can't be biased because it's AI"*	S2-P110
Q.21	*"[...] it may be fair as an algorithm does not take into account factors such as someone's manner or dress which may lead to an unconscious bias for or against an applicant when assessed by a human."*	S2-P8
Q.22	*"It is very fair because all applicants are assessed using the same list of criteria."*	S2-P85
Q.23	*"It takes in essential information needed to evaluate weather a loan is risky from the bank's point of view as a business deal, it doesn't take feelings or emotions, just facts, and applies them to the bank's set criteria with which they are happy to give a loan out to."*	S2-P98
Q.24	*"I think they have asked the correct information to see if an individual could be able to afford to pay back the loan."*	S2-P34
Q.25	*"I think it is fair that it is based on the same factors for everyone but there are circumstances under which more personal information individual to their case should be taken into consideration."*	S2-P51
Q.26	*"The AI system will only deal with data/numbers and won't take into consideration Kim's personal circumstances which could explain why she was rejected in the first place. For example, many lost their jobs due to no fault of their own during the pandemic and fell behind on bills etc. and many have ended up in debt. If this was the case with Kim it wouldn't really be fair based on the circumstances."*	S2-P96

Q.27	*"Everyone is treated the same, but it seems that if a human saw she was only 5% off having the loan, they would have just let it slide."*	S2-P9
Q.28	*"There should be some human to evaluate those cases that are in the obscure region of the cutting-off point."*	S2-P209
Q.29	*"If the person trying to get the loan is rejected within a small margin and appeals I believe they should be able to re-negotiate."*	S2-P185
Q.30	*"They took the human element away, which allows for communication and some compromise."*	S2-P245
Q.31	*"[...] there will always be instances where an AI will get the decision wrong when a person land in a grey area/their circumstances fall into an area where a little compassion is needed."*	S2-P218

Table 4: Summary of some of our participants' responses to the open ended questions. S1 = preliminary study, S2 = main study, Pj = index of the participant.

B SUMMARY OF THE EXPERIMENTAL DESIGN

Parameters	Conditions	Descriptions
Explanation	No explanation	*The artificial intelligence system uses some of this information for making the loan decision.*
	With explanations	*In the email received by Kim, an explanation of how the decision-making system has reached the conclusion is included. The email includes the importance that each piece of information provided by Kim had in the final decision. Factors are listed from the most important to the least important factor based on the bank's criteria. The magnitude of the contribution of each piece of information (negative (−) means that it contributed to the rejection decision) is added between brackets:* *Credit Score* (−0.15) > *Loan amount requested* (−0.12)> *Total annual income* (−0.09)> *Loan purpose* (+0.02)> *Employment status* (+0.02)> *Loan amount term (months)* (−0.03)> *Date of birth* (+0.03)> *Co-applicant (if any) income* (+0.01)> *Number of dependents* (−0.07)> *Education* (+0.02) *The email also includes information about scenarios where the individual would have been granted the loan. Kim would have been granted a loan if one of the following scenarios had been true:* • *The loan amount requested had been* 5% *lower* • *The total annual income of the individual had been* 10% *higher* • *The credit score of the individual had been "Very Good"*
Human oversight	No human oversight	*Given the latest technological advances and in an effort to make loan decisions in a timely manner, the loan application process is now fully automated. An artificial intelligence system receives the online requests and evaluates each case. An email is sent to the applicants with the final verdict.*
	With human oversight	*Given the latest technological advances and in an effort to make loan decisions in a timely manner, the loan application process is now hybrid: it combines artificial intelligence with human expertise. This involves a two-step approval process. In the first step, an artificial intelligence system receives the online requests and evaluates each case. If the artificial intelligence system reaches a decision (approve or reject) with a high confidence, an email is sent to the applicant with the final verdict. If the artificial intelligence system has a low confidence over the decision, there is a second step where a human oversees the decision and makes the final verdict and an email is sent to the applicant.*
Contestability	No contestability	*Since the reason for introducing an artificial intelligence system is to handle home loan applications in a timely manner, Kim has no option to request a review of the decision.*
	Contest initial decision	*Kim has decided to appeal the decision and has asked for a review of the process. As part of the review procedure, Kim has the opportunity to make objections about the initial decision and provide any information to support the application. The same artificial intelligence system will then reevaluate the home loan application.*
	Contest decision maker	*Kim has decided to appeal the decision and has asked for a review of the process. As part of the review procedure, Kim has the opportunity to ask for a human to review the process. This human reviewer will make a completely new decision with the information that Kim already provided for the initial decision.*
Task stakes	High stakes	*Buy a house / home loan*
	Low stakes	*Go on holiday / holiday loan*

Table 5: Summary of the experimental design.